PRICE INTELLIGENCE (PI) – DATA GATHERING BY UTILIZING WEB CRAWLING

Big Data Applications and Utilising Non-Traditional Data Sources and Methods for Official Statistics

10 June 2021

Ms. Mazliana Mustapa
Department of Statistics Malaysia
The Department of Statistics Malaysia (DOSM) has initiated the implementation of BDA under the project of Statistics Big Data Analytics (STATSBDA).
STATSBDA OBJECTIVE

- To modernise data collection
- To allow high accuracy
- To reduce respondents burden
- To use as supplements for existing data in production of certain statistics
- To produce new statistical indicators

Objective
PRICE INTELLIGENCE (PI)

Price Intelligence

- leveraging the capability of Big Data in collecting large data from various sources and transform them into better structure

- different prices of the same good can be obtained through various online retailer websites, providing a modernized price data collection

- Transform from unstructured data into structured data to perform analysis

What is web crawler?

process of repetitively finding and fetching hyperlinks starting from a list of starting URLs.

What is web scraping?

Web-scraping is automatically retrieving and processing information from websites
PRICE INTELLIGENCE OBJECTIVE

Objective

01
to give better insight in consumer price analysis and monitoring

02
to establish new price basket analysis which will be used as value added to the current Consumer Price Index

New Data Collection Methodology
New data collection process introduced to cover price of goods selling online

Value Creation

Holistic View in Online and Offline Prices
Allow monitoring and forecast future price trend and as valuable input for price control decisions by government

Transform of Business Process
New data collected enable to create holistic landscape of CPI monitoring process

PI Visualization
Data Management
<table>
<thead>
<tr>
<th>DATA ACQUISITION</th>
<th>ETL</th>
<th>PRICE LAKE DEVELOPMENT</th>
<th>ANALYSIS</th>
<th>VISUALIZATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>External Data Sources</td>
<td>Online Retail Database</td>
<td>Raw Data Management</td>
<td>Trending &amp; Forecast</td>
<td>Dashboard</td>
</tr>
<tr>
<td>Online Retailer</td>
<td>Intelligent Product Classification (COICOP)</td>
<td></td>
<td>Descriptive</td>
<td></td>
</tr>
<tr>
<td>Website Government/Agency</td>
<td>FTP</td>
<td>Enrichment Data Management</td>
<td>Dependency &amp; Correlation</td>
<td></td>
</tr>
<tr>
<td>DOSM Price Database (CPI)</td>
<td></td>
<td></td>
<td>Price Basket Enrichment</td>
<td></td>
</tr>
</tbody>
</table>
MODULE IN PRICE INTELLIGENCE

1. **Data Acquisition**
   - Collect & Extract data from various sources, both internal & external

2. **Price Lake Development**
   - Store & structure massive data

3. **Data Analysis**
   - Descriptive
   - Exploration
   - Comparison
   - Dependency

4. **Data Visualization**
   - Make sense of data

Data Classification (COICOP dictionary)

Note: COICOP → Classification of Individual Consumption According to Purpose
DATA MANAGEMENT

PI Dictionary

The word most often appear on the item description

Pending/Accepted refers to the matching result at the 4 Digit COICOP level

Partially/exact refers to the matching result with MCOICOP

11.3% exact match
88.7% partially match

*Information based on 231 item specifications updates in PI Dictionary
**DATA MANAGEMENT**

**PI Dictionary – Text matching scoring**

Specification that meet requirement of items in CPI basket of goods with high scoring value will return exact match, mapped to 7 digits item specification MCOICOP code

<table>
<thead>
<tr>
<th>Item Specification</th>
<th>6D Code</th>
<th>Item Code</th>
<th>Score</th>
<th>Match Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. huggies dry diapers m 6-11 kg 72 pieces each</td>
<td>121321</td>
<td>1213212</td>
<td>75.14</td>
<td>EXACT</td>
</tr>
<tr>
<td>i. drypers wee wee dry disposable diapers m 6-11kg 74pcs each</td>
<td>121321</td>
<td>1213211</td>
<td>76.59</td>
<td>EXACT</td>
</tr>
<tr>
<td>i. huggies dry diapers m 6-11 kg 72 pieces each</td>
<td>121321</td>
<td>1213212</td>
<td>75.14</td>
<td>EXACT</td>
</tr>
<tr>
<td>i. huggies dry diapers m 6-11 kg 72 pieces each</td>
<td>121321</td>
<td>1213212</td>
<td>75.14</td>
<td>EXACT</td>
</tr>
<tr>
<td>i. drypers wee wee dry disposable diapers m 6-11kg 74pcs each</td>
<td>121321</td>
<td>1213211</td>
<td>76.59</td>
<td>EXACT</td>
</tr>
<tr>
<td>i. drypers wee wee dry disposable diapers m 6-11kg 74pcs each</td>
<td>121321</td>
<td>1213211</td>
<td>76.59</td>
<td>EXACT</td>
</tr>
</tbody>
</table>

Meanwhile, broad specification items has only 6 digits item code to be matched with online item specification. Items is considered matching as long as the item is the same regardless of brand name, units etc

<table>
<thead>
<tr>
<th>Item Specification</th>
<th>6D Code</th>
<th>6D Code</th>
<th>Item Code</th>
<th>Score</th>
<th>Match Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. drypers wee wee dry disposable diapers xol 15+kg 40pcs each</td>
<td>121321</td>
<td></td>
<td></td>
<td>67.12</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>i. huggies dry pants diapers l 9-14kg 50pcs each</td>
<td>121321</td>
<td></td>
<td></td>
<td>64.44</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>i. tena value diapers, medium 8 pack x 12s</td>
<td>121321</td>
<td></td>
<td></td>
<td>50.86</td>
<td>PARTIAL</td>
</tr>
</tbody>
</table>
PI Dictionary has been built in order to get the best match to MCOICOP 6/7 digits based on description of the items using text matching scoring.

Dictionary will be updated regularly based on basket of goods in CPI.

**PI Dictionary**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>4D Code</th>
<th>AD Code</th>
<th>Description</th>
<th>Sample Keyword</th>
<th>Sample Variation</th>
<th>Item Code</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>wearables</td>
<td>1213</td>
<td>121321</td>
<td>Lampon pakai buang untuk bayi/ Diapers for baby</td>
<td>Baby Diapers, Disposable Diapers, Drypers, Gerbi, Gote, Huggles, Manyokko, Pampers, Whoopec</td>
<td>Classic, Diapers, Drynings</td>
<td>1213211</td>
<td>✔️</td>
</tr>
<tr>
<td>wearables</td>
<td>1213</td>
<td>121321</td>
<td>Lampon pakai buang untuk bayi/ Diapers for baby</td>
<td>Baby Diapers, Disposable Diapers, Gerbi, Gote, Huggles, Lampon Pakai Buang Sayi, Manyokko, Pampers, Tesco Loves, Whoopec</td>
<td>Baby, Diapers, Dry, Drynings, Drygoods, Extra, Pants, Tape, Wise Wise Dry</td>
<td>1213212</td>
<td>✔️</td>
</tr>
<tr>
<td>wearables</td>
<td>1213</td>
<td>121322</td>
<td>Syampu/perai/ Shampoo/conditioner</td>
<td>COCONI, Clairol, Pantene, Perai, Shampoo, SHIP, SHIP, SOK</td>
<td>Clairal, Shampoo, Pantene</td>
<td>1213223</td>
<td>✔️</td>
</tr>
</tbody>
</table>

**COICOP**: Classification of Individual Consumption According to Purpose
1. Click Data Monitoring on the top left of the screen.
2. Click Folder Name and Click file name.
3. Status of the generated data will be displayed. Click Download to download to local PC.
b) Descriptive
Descriptive analysis, more emphasis on data exploration summaries such as: Mean, Median, Standard Deviation, Variants, Histogram / Skewness, etc.

a) Trend
Trend analysis is used to evaluate data patterns based on linear approaches. There are various time series analysis techniques that can be used, such as ARIMA, Exponential Smoothing, Holt Winter, Linear Trend, Exponential and Level Aggregation.
c) Price Basket Enrichment

Price Basket Enrichment is the process of adding data crawled from online sources into available data from the SIHP-CPI System. This process requires DOSM support because not all prices used in CPI manual calculation are available online.

d) Dependency

Dependency analysis looks for more connections between data. An example is to evaluate variables by performing factor analysis or releasing irrelevant data. Cluster analysis can be carried out to assess data relationships (correlations) and other analysis can be conducted to assess such as Factor Analysis - Cause Analysis, Correlation and Cross Category.
During the pandemic Covid-19, online flight ticket prices has been used in the compilation of the CPI.

Item specification (Route)

- Kedah to Kuala Lumpur
- Johor Bahru to Kuala Lumpur
- Kuantan to Kuala Lumpur
- Penang to Kuala Lumpur
- Kuala Terengganu to Kuala Lumpur

Implications of the Movement Control Order (MCO), all the price data collection at the outlet has been suspended. Data crawling has been done during the MCO for 20 main product CPI as below:

- CAP RAMBUTAN HIJAU SST RICE 5% 10KG (RICE)
- JATI BERAS SUPER SPESIAL 10KG (RICE)
- JASMINE RICE SUPER SPECIAL 5% 10KG (RICE)
- JASMINE RICE SUPER SPECIAL TEMPATAN 5KG (RICE)
- AYAM BERSIH (PELBAGAI BAHAGIAN) (CHICKEN)
- IKAN BAWAL HITAM (FISH)
- IKAN CENCARU (FISH)
- IKAN KEMBUNG (FISH)
- KUETIAU BASAH (FLAT RICE NOODLES)
- MEE KUNING BASAH (NOODLES)
- UBI KENTANG (POTATO)
- BAWANG BESAR (ONION)
- CILI KERING KERINTING (DRIED CHILLI)
- KACANG BUNCIS (FRENCH BEAN)
- KUBIS BULAT (TEMPATAN) (CABBAGE)
- LOBAK MERAH (CARROT)
- TOMATO
- SANTAN KELAPA (FRESH COCONUT MILK)
- TELUR AYAM GRED B (HEN'S EGGS GRADE B)
- *MINYAK MASAK (PELBAGAI JENAMA) (COOKING OIL)
APPLICATIONS OF PRICE INTELLIGENCE

HOUSE RENTAL PRICES FOR INTERNATIONAL COMPARISON PROGRAM (ICP)

45.6% housing rental data for International Comparison Program 2017 (ICP2017) submission were using online price data

RENTAL HOUSING DATA SOURCED USED FOR ICP 2017

19 types of housing specification with different sizes consisting of
• Single-detached house
• Attached house (row house)
• Studio apartment
• One-bedroom apartment
• Two-bedroom apartment
• Three-bedroom apartment
LIMITATION

- Need to update crawler
- Have to build a bunch of crawlers for different sites
- The structure of websites change frequently
- Legal issues involved
- Storage limitation (huge amounts of data)
- Access and scrape data which is publicly available and avoid trying to crawl data which is private or protected by copyrights and other laws;
- Always check the website’s robots.txt file
LEGALISATION

Intellectual Property
The content of the Tesco Website is protected by copyright, trade marks, database and other intellectual property rights. You may retrieve and display the content of the Tesco Website on a computer screen, store such content in electronic form on disk (but not any server or other storage device connected to a network) or print one copy of such content for your own personal, non-commercial use, provided you keep intact all and any copyright and proprietary notices. You may not otherwise reproduce, modify, copy or distribute or use for commercial purposes any of the materials or content on the Tesco Website without written permission from Tesco Website.

No licence is granted to you in these Terms and Conditions to use any of our trade marks.
A robots.txt file tells search engine crawlers which pages or files the crawler can or can’t request from your site. This is used mainly to avoid overloading your site with requests; it is not a mechanism for keeping a web page out of Google. To keep a web page out of Google, you should use noindex directives, or password-protect your page.

A robots.txt file is used primarily to manage crawler traffic to your site, and usually to keep a file off Google, depending on the file type.

Source: https://developers.google.com/search/docs/advanced/robots/intro
1. Forecasting prices of fish and vegetable using web scraped price micro data  
   - Mazliana Mustapa, Raja Rajeswari Ponnusamy, Ho Ming Kang

2. Online and Offline Prices: Measuring selected home appliance’s products  
   - Mohd Saiful Husain and Norsyela Muhammad Noor Mathivanan

3. Analysis of the Mobile Phones Prices Malaysia using Web Scraped Data  
   - Nur Hurriyatul Huda Abdullah Sani

Table 3: ARIMA best model

<table>
<thead>
<tr>
<th>Item</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Bream</td>
<td>$Y_t = 2.4205 -0.0333Y_{t-1} + 0.6382Y_{t-2} -0.0594 + 0.9688e_{t-1}$</td>
</tr>
<tr>
<td>Solar Kuning</td>
<td>$Y_t = 1.4794 - 1.6648Y_{t-1} - 0.7431Y_{t-2} + 0.7924e_{t-1}$</td>
</tr>
<tr>
<td>Green Spinach</td>
<td>$Y_t = 1.1813 + 1.3246Y_{t-1} - 0.7774Y_{t-2} + 0.1088 - 0.8126e_{t-1}$</td>
</tr>
<tr>
<td>Kangkung</td>
<td>$Y_t = 1.0995 + 1.7293Y_{t-1} - 0.7776Y_{t-2} + 0.1023 - 0.8315e_{t-1}$</td>
</tr>
<tr>
<td>Long Beans</td>
<td>$Y_t = 0.5460 + 0.3667Y_{t-1} + 0.0780 + 0.4322e_{t-1} + 0.2638e_{t-2}$</td>
</tr>
<tr>
<td>Bawal, Cencur, Kembong,</td>
<td>$Y_{t, \mu} + Y_{t, \lambda}$</td>
</tr>
<tr>
<td>Round Cabbage and Sawi</td>
<td>where: $\mu$ mean of the changes of period to period</td>
</tr>
<tr>
<td>Jepun</td>
<td></td>
</tr>
</tbody>
</table>
CONTACT INFO

Mazliana Mustapa
Core Team Big Data Analytics
Department of Statistics Malaysia

Team Price Intelligence
mazliana@dosm.gov.my
ridhuan@dosm.gov.my
noradilah.adnan@dosm.gov.my

https://statsbda.dosm.gov.my/
THANK YOU

BANCI MALAYSIA